Nama / NIM: Ilham Wahyu Adli / 1301173380

Kelas: IF-GAB-02

---

1. Sepuluh unigram yang paling sering muncul

```
:
# Soal 1
newFreqUni = freqUni.copy()
del newFreqUni['<s>']
del newFreqUni['</s>']
sort_orders = sorted(newFreqUni.items(), key=lambda x: x[1], reverse=True)[:10]
for j in (sort_orders):
    print(j[0], j[1])
```

```
yang 359
di 291
staycation 255
dan 190
hotel 186
bisa 162
dengan 159
untuk 144
ini 108
kamu 107
```

2. Sepuluh bigram dengan probability paling tinggi

```
# Soal 2
newProbBi = probBi.copy()
sort_orders = sorted(newProbBi.items(), key=lambda x: x[1], reverse=True)[:10]
for j in (sort_orders):
    print(j[0], j[1])
```

```
('paket', 'staycation') 1.0
('kendati', 'demikian') 1.0
('diketahui', 'sebagian') 1.0
('mengutip', 'how') 1.0
('how', 'stuff') 1.0
('stuff', 'works') 1.0
('works', 'pengertian') 1.0
('pengertian', 'staycation') 1.0
('kombinasi', 'dari') 1.0
('membuatnya', 'menarik') 1.0
```

3. Kalimat Uji yang digunakan adalah:

Sesuai dengan topik

- *Staycation* adalah cara liburan yang cenderung hemat
  Terdapat beberapa artikel yang menjelaskan jika *Staycation* adalah liburan yang terhitung hemat.
- *Staycation* adalah kegiatan berlibur di dekat rumah
  Artikel dominan menjelaskan pengertian ap aitu *Staycation*.
- *Staycation* alternatif liburan yang paling diminati warga Jakarta
  Dalam 3 artikel terakhir terdapat rekomendasi hotel bagi warga Jakarta.

Tidak sesuai dengan topik

- Facebook merupakan sosial media dengan pengguna tertinggi di dunia
- Pembelian sepeda meningkat pesat pada saat pandemi COVID
- PSBB merupakan salah satu jalan yang ditempuh pemerintah Indonesia

4. Analisis perplexity, perbandingan antara hasil yang diperoleh dari kalimat dengan topik yang mirip dan yang tidak mirip.

```
['<s>', 'staycation', 'adalah', 'cara', 'liburan', 'yang', 'cenderung', 'hemat', '</s>']
Probability:  2.4960142868071624e-30
Perplexity:  5016
None
['<s>', 'staycation', 'adalah', 'kegiatan', 'berlibur', 'di', 'dekat', 'rumah', '</s>']
Probability:  9.681338917172998e-29
Perplexity:  3175
None
['<s>', 'staycation', 'alternatif', 'liburan', 'yang', 'paling', 'diminati', 'warga', 'jakarta', '</s>']
Probability:  1.5188927096970402e-35
Perplexity:  7391
None
['<s>', 'facebook', 'merupakan', 'sosial', 'media', 'dengan', 'pengguna', 'tertinggi', 'di', 'dunia', '</s>']
Probability:  4.21515892602595e-43
Perplexity:  17279
None
['<s>', 'pembelian', 'sepeda', 'meningkat', 'pesat', 'pada', 'saat', 'pandemi', 'covid', '</s>']
Probability:  7.287167849307812e-39
Perplexity:  17278
None
['<s>', 'psbb', 'merupakan', 'salah', 'satu', 'jalan', 'yang', 'ditempuh', 'pemerintah', 'indonesia', '</s>']
Probability:  1.856818027662577e-41
Perplexity:  11834
None
```

Dapat dilihat dari hasil program yang telah dibangun, jika kita melakukan percobaan menggunakan data uji dengan topik yang sama yaitu "*staycation*" maka hasilnya memiliki probabilitas yang lebih besar walaupun kalimatnya lebih panjang seperti pada contoh kalimat ke-3. Dimana dalam kalimat ke-3 memiliki jumlah kata yang sama dengan kalimat ke-5 tetapi nilai probabilitynya lebih besar (dapat dilihat dari e-xx). Dan seluruh nilai perplexity dari kalimat yang mengandung unsur "*staycation*" memiliki angka perplexity yang jauh lebih rendah dibandingkan dari kata yang tidak termasuk kedalam topik "*staycation*".